DOCUMENT RESUME

ED 274 714 TM 860 593

AUTHOR Olsen, James B.; And Others
TITLE Comparison and Equating of Paper-Administered, Computer-Administered and Computerized Adaptive Tests of Achievement.
PUB DATE Apr 86
NOTE 46p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Tests; *Adaptive Testing; *Comparative Testing; *Computer Assisted Testing; Elementary Education; Equated Scores; Grade 3; Grade 6; Item Analysis; Latent Trait Theory; Scores; *Test Format; Test Length
IDENTIFIERS California Assessment Program; *Paper and Pencil Tests; PLATO

ABSTRACT

Student achievement test scores were compared and equated, using three different testing methods: paper-administered, computer-administered, and computerized adaptive testing. The tests were developed from third and sixth grade mathematics item banks of the California Assessment Program. The paper and the computer-administered tests were identical in item content, format, and sequence. The computerized adaptive test consisted of an adaptive subset of the computer-administered test. A total of 575 students were administered two of the three types of tests in varying sequence. PLATO/WICAT System 300 computers were used to administer the tests. Although the paper and computer-administered tests contained 55 or 62 items, the adaptive test usually contained twenty items. Item response theory and analysis of variance were performed. Results supported the comparability of paper-administered and computer-administered tests. The computer-administered test required half to three-quarters as much testing time as the paper and pencil test; the adaptive test required one-fourth as much time as the paper and pencil test. Scores tended to be lower on the second test administered. Ability estimates for all three tests were highly correlated. Both the two- and the three-parameter item response theory model were suitable. Three pages of references are included. The two appendixes consist of: (1) comparisons of item parameters; and (2) test information and standard error curves. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED274714

COMPARISON AND EQUATING OF PAPER-ADMINISTERED, COMPUTER-ADMINISTERED AND COMPUTERIZED ADAPTIVE TESTS OF ACHIEVEMENT

James B. Olsen, Dennis D. Maynes, Dean Slawson, and Kevin Ho

Waterford Testing Center

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. B. Olsen

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM 860 593

Paper presented at the American Educational Research Association Meeting
San Francisco, April 1986.

BEST COPY AVAILABLE

PURPOSE

This research study was designed to compare and equate student achievement scores from three different testing methods: paper-administered testing, computer-administered testing, and computerized adaptive testing. All three of the testing formats were developed from the California Assessment Program (CAP) item banks for grades 3 and 6. The paper-administered and the computer-administered tests were identical in item content, format, and sequence. The computerized adaptive test consisted of an adaptive subset of the items in the computer-administered test.

INTRODUCTION

Paper-Administered Testing.

Since the early 1900's the fixed length, paper-and-pencil test has been the standard method for administering student achievement or assessment tests. These standardized achievement tests are generally administered by schools only once or twice a year. The students within each grade respond to a fixed set of items selected at the appropriate grade level. Some of these items are often too easy and some are often too difficult for any given student. In the majority of achievement tests all students take the same fixed set of items. However, assessment alternatives have been developed to reduce the number of test items to be administered to any given student. These include random selection of items from pre-specified domains, matrix sampling, and multiple matrix sampling; with these models each student receives a different set of test items. For the majority of standardized achievement and assessment tests, the primary item selection method used is to select items which measure well students of average levels of ability. Research has shown that this testing procedure necessarily results in less than optimal measurement for the majority of students and significantly poorer measurement of achievement for students at very low and very high ability levels. (Weiss, 1975, Lord, 1980; Hambleton and Swaminathan, 1985). With many of the matrix sampling assessment procedures, only group or school level reports are available; no individual student reports are provided.

Several studies have also shown that standardized, paper-and-pencil achievement tests are of limited use in evaluating the impact or change due to educational programs and curriculum. (Armbruster, Stevens, and Rosenshine, 1977; Floden, Porter, Schmidt and Freeman, 1977; Freeman, Kuhs, Knappen, and Porter,1979; House, Glass, McLean and Walker, 1978; Tyler and White, 1979).

A further limitation of traditional paper-administered achievement tests is the extensive time delay between the testing date and the date when the results are reported. Due to the above reasons, traditional paper-administered achievement tests are seen to be of limited use in helping teachers and students in the daily educational tasks (Tyler and White, 1979).

Computer-Administered Testing.

Recent developments in microcomputer technologies have demonstrated the potential for microcomputers to provide computerized test administration, scoring, and reporting. Computerized test administration offers the following advantages over printed test administration: more standardized

test administration, improved test security, individualized testing, immediate test scoring and feedback, increased variety of testing formats, and ability to collect test and item latency information. Advanced microcomputers also offer the capability of prescription generation and assignment of computerized lessons to remediate deficiencies.

Computerized Adaptive Testing and Item Response Theory.

Concurrent developments in item response theory (IRT) and microcomputer technology have spurred significant research and development of computerized adaptive testing (CAT) or tailored testing systems and procedures (Lord, 1980, McBride, 1977,1981; Olsen, Bunderson, and Gilstrap, 1982,1983; Weiss, 1979,1981,1983). In an adaptive test, the test items selected for administration are based on the examinee's responses to previously administered items.

In developing a computerized adaptive test a large pool of test items is written and administered to a large representative sample of students. From the representative sample of item responses, the test items are calibrated with parameters indicating the difficulty, discrimination and guessing levels for the items. Once the item pool is calibrated a computerized adaptive test involves making an initial ability estimate and selecting an item from the pool for presentation to the examinee. If the student answers the item correctly a more difficult item is presented; if the student misses the item a less difficult item is presented. After each response the ability estimate is updated and a new item is selected at the new estimated ability level. The testing process continues until a specified level of reliability or precision is reached and the testing process is terminated. Immediately, a report can be provided to the teacher and/or the student which indicates the current ability level of the student.

Computerized adaptive tests have been shown to take less than half of the testing time as traditional achievement tests and to provide more precise ability estimates across the entire ability range. The ability estimates and the item parameters are calibrated on a common scale so that these estimates are theoretically independent of the particular sample of persons taking the test and the particular sample of items taken by each examinee. This allows for accurate estimation of ability on the common scale even though different groups of items are taken by different examinees.

Computerized adaptive tests are generally based on the assumption of a unidimensional latent trait or ability which characterizes all items in the subtest bank. Recent developments in multivariate latent trait theory offer the future possiblity for implementation of multidimensional item response models. (Reckase, 1985; Hambleton and Rovinelli, 1983; Whitely, 1980; Samejima, 1974).

The theory and applications of computerized adaptive testing have primarily been applied to the estimation of ability tests rather than achievement tests. However, significant applications of IRT theory to the measurement of student achievement have been demonstrated (Bejar, Weiss and Kingsbury, 1977; Brown and Weiss, 1977; Hambleton, 1979, 1983; Kingsbury and Weiss, 1981, 1983; Schnitz, Maynes and Olsen, 1983; Traub and Wolfe, 1982; Weiss, 1979, 1981, 1983 and Weiss and Kingsbury, 1984).

In summary, the present study is one of the first to compare and equate achievement scores from three alternative methods of testing delivery: paper-administration, computer-administration, and computerized adaptive test administration. This is also a pioneering study to evaluate computerized adaptive testing at the elementary grade school level.

METHODS

This study compares three different modes of test administration: paper-administered tests, computer-administered, and computerized adaptive tests. These three alternative testing formats were prepared from the mathematics application items in the California Assessment Program (CAP) item bank. The paper-administered tests and the computer-administered tests were sequentially administered and included 55 items at grade 3 and 62 items at grade 6. The computerized adaptive test was administered using adaptive testing procedures and included a stopping rule after 20 items or a posterior standard deviation of .05, whichever came first. The computer-administered and computerized adaptive tests were delivered on a PLATO/WICAT System 300 school microcomputer system with thirty terminals.

Students at each grade (3 and 6) were randomly assigned to one of four experimental groups. Each group received two different test formats according to the following design.

	First Test		Second Test
Group 1:	Computer Administered Test	--	Computerized Adaptive Test
Group 2:	Computerized Adaptive Test	--	Computer Administered Test
Group 3:	Paper Administered Test	--	Computerized Adaptive Test
Group 4:	Computerized Adaptive Test	--	Paper Administered Test

All students received general test directions followed by directions for recording their answers on the answer sheet or the computer keyboard. The students then completed the designated tests in the order presented above for their test group.

SUBJECTS

Subjects for the study were selected for participation from three California school districts and one or two schools per district. The school districts selected for participation were Port Hueneme, Moorpark, and Santa Barbara. These school districts were selected because they had already installed the PLATO/WICAT System 300 school computers and were interested in participation in the study. At each of the schools, all students in grades 3 and 6 participated in the research study. A total of 350 students at grade 3 and 225 at grade 6 were involved in the study. Tests were administered in the schools between May 20 and June 5, 1985.

PROCEDURES

The paper-administered tests and the computer-administered tests each consisted of 55 mathematics application items at grade 3 and 62 mathematics application items at grade 6. The computerized adaptive test consisted of 15-20 items selected from the grade level mathematics applications item bank according to computerized adaptive testing principles. For the computerized adaptive test, the test exit criterion was twenty items or a posterior standard deviation of .05 or smaller, whichever came first. Research has shown that twenty items in a computerized adaptive test will provide a high level of measurement precision. (Stocking, 1984).

The research design presented above was executed through a computerized program which randomy assigned each new student per grade to one of the four research groups. The computer program then displayed to the student the first selected test according to the research design. If the first selected test was a computer-administered test or a computerized adaptive test, then the program sequenced the student to the first computerized test. After completing the first test, the program sequenced the student directly to the second test specified by the design. If the first selected test was a paper-administered test, the program told the student to request a specific color-coded form from the test proctor (blue for grade 3 and yellow for grade 6).

Each student was given a unique computer login name to distinguish the student from all other students. Student login names were linked to both the paper-administered tests, the computer-administered tests, and the computerized adpative tests.

INDEPENDENT AND DEPENDENT VARIABLES

The independent variables for this study were test administration method (paper-administered, computer-administered, and computerized adaptive), grade level (3 and 6), and order of test presentation. The dependent variables were student achievement scale scores, and testing time.

COMPUTERIZED ADAPTIVE TESTING PROCEDURES.

Computerized adaptive testing involves four major steps. The first step is to select an initial starting point or initial ability estimate for the examinee. The starting point used in this study was an item drawn from the average level of difficulty (b=0.0). The second step is to calculate or update the ability estimate based on the examinee's response and item parameter values. Three major methods are available for ability estimation (sequential Bayesian, EAP, and maximum likelihood). This study used the EAP ability estimation algorithm (Bock and Aitkin, 1981; and Bock and Mislevy, 1982). The third step is to select the next item given the updated ability estimate. An item information matrix was used for item selection. The item information matrix included 20 levels of item difficulty and rank ordered 20 items within each of the 20 difficulty levels in terms of decreasing item information. Based on the new updated ability estimate, a random selection was made of one of the four most informative items not yet administered at the designated difficulty level. The new item was administered to the examinee and a new ability estimate calculated. Step four involves continuing the computerized adaptive testing process until a designated stopping rule is met. The stopping rule selected for this study was a minimum standard error (posterior standard deviation) of less than .05 or a maximum test length of 20 items.

DATA COLLECTION AND ANALYSIS

The computer scan sheets from the paper-administered tests were scanned, scored and transferred to the master computer data files. Computer

tapes containing results from the computer-administered and computerized adaptive tests were collected from each participating school. From the computer tapes, item response files were prepared for each administration method, test order and grade level.

Since the paper-administered tests and the computer-administered tests were scored with the number right metric and the computerized adaptive tests were scored with the ability (theta) metric, there was a need to convert all of the item scores and test scores to the same common ability scale metric.

The BILOG item calibration program (Mislevy and Bock, 1982) was used to calculate item parameters simultaneously for the paper-administered items, the computer-administered items, and the computerized adaptive-administered items. This was accomplished by aggregating the data sets for each method of test administration and calculating the number of examinees trying each item and the number of examinees correctly answering for each item across all test administration modes. This allowed the use of the largest combined data set for full calibration. The item parameters for each of the test formats were thus scaled on the same common scale. From the item parameter estimates and the item response vectors, ability estimates were calculated for each of the three methods of test administration. Since each student received two different test formats, two ability estimates were calculated and available for each student according to the research design.

The BILOG item calibrations were calculated using one, two, and three parameter item response theory models. Comparisons were made between the item fit statistics calculated from the three different item parameter models. Test information and standard error curves were estimated for each of the three models.

RESULTS

Analysis of Variance.

A repeated measures analysis of variance was calculated between the total scores the the paper-administered tests and the computer-administered tests. This analysis adjusts for individual student variation by examining deviations from the individual's overall average performance. The results of this analysis are presented in Tables 1A and B.

Table 1A
Analysis of Variance Results for
Paper and Computer Administered Tests
Grade 3

Source	Df	Sum of Squares	Mean Square	F-Ratio	P-Level
Test Mode	1	.132514	.132514	.015382	.901
Order	1	24.904044	24.904044	2.810922	.095
T X O	1	.062961	.062961	.007084	.933
Error	210	1866.400481	8.887621		

Table 1B
Analysis of Variance Results for
Paper and Computer Administered Tests
Grade 6

Source	Df	Sum of Squares	Mean Square	F-Ratio	P-Level
Test Mode	1	.002320	.002320	.000259	.987
Order	1	42.831724	42.831724	4.843929	.030*
T X O	1	.002431	.002431	.000275	.987
Error	125	1104.330223	8.834642		

* $p < .05$

For Grade 3, Table 1A shows no significant Test Mode, Order or Test x Order interactions. For Grade 6, there was a significant Order effect but no significant Test x Order interaction. These results indicate a high degree of comparability between the total raw scores for paper-administered tests and for computer-administered tests. Table 2 presents a table of the descriptive statistics for each main effect.

Table 2

Descriptive Statistics for Test Mode and Order

	Test Mode		Order	
	Paper	Computer	First	Second
Grade 3				
MEAN	-.03242	.01871	.33872	-.35244
SD	.29888	.28392	.28810	.29486
N	102	112	108	106
Grade 6				
MEAN	.00013	.00856	.58078	-.57209
SD	.37172	.36906	.37172	.36906
N	64	65	64	65

The calculated means in Table 2 are average deviations of student performance from their overall average performance.

Table 2 shows no differences between paper-administered tests and computer-administered tests for Grades 3 and 6. The means for test order effects show that the second test administered has lower scores (.68 to 1.05 points) than the the first test administered. This could be due to a test fatigue effect or other reasons.

Item Analysis.

Item analyses were conducted for the paper-administered tests and the computer-administered tests. The descriptive statistics from the item analyses using raw scores are presented in Table 3. There were no significant differences between the means or score distributions for the computer-administered and paper-administered tests. Alpha reliabilities and the standard errors of measurement are very similar for the computer-administered and paper-administered tests.

Table 3
Test Statistics for Mathematics Application Items
Administered by Computer or Paper

	MEAN	MED	SD	N	RANGE	RELIAB	SEM
Grade 3, 55 items total							
CA	36.47	37.56	11.647	263	49	.924	3.211
PA	35.76	36.33	12.336	214	52	.930	3.265
Grade 6, 62 items total							
CA	30.47	29.88	12.227	171	51	.909	3.684
PA	30.49	28.67	12.196	129	48	.910	3.660

KEY

CA = Computer-Administered
PA = Paper-Administered
MED = Median
RELIAB = Alpha Reliability
SEM = Standard Error of Measurement

Testing Time.

Table 4 shows the testing time in minutes required for the paper-administered, the computer-administered test and the computerized adaptive test. This table shows significant differences between the testing time required for the different testing formats. For grade 3, the computerized adaptive test took one-fourth of the time of the paper-administered test; the computer-administered test took half of the time of the paper-administered test. At grade 6, the computerized adaptive test took one-fourth of the time of the paper-administered test; the computer-administered test took three-fourths of the time of the paper-administered test.

Table 4
Testing Time in Minutes Required By Paper-Administered, Computer-Administered, and Computerized Adaptive Tests

		PA	CA	CAT
Grade 3	MEAN	45.01	25.73	10.29
	SD	19.74	9.03	5.04
	N	172	264	266
Grade 6	MEAN	54.03	41.95	14.41
	SD	19.45	15.16	5.56
	N	99	173	184

Ability Estimates.

Ability estimates based on a three parameter IRT model were calculated for the three methods of test administration. A total ability estimate was also calculated based on the total number of items tried and the total number of items correct across the methods of test administration for each given student. Table 5 presents the descriptive statistics for the ability estimates for each grade and testing method. These data show very similar and approximately normal ability distributions for each method of test administration. The computer-administered and computerized adaptive tests show slightly higher means and slightly lower standard deviations than the paper-administered tests. The same pattern is found in both grades 3 and 6. The ability estimates are approximately normally distributed with a mean of 0.0 and a standard deviation of 1.0. An ability estimate of 0.0 would correspond approximately to the 50th percentile.

Table 5
Descriptive Statistics For Ability Distributions
And Different Testing Formats

	Mean	Variance	SD	N
GRADE 3, MATH				
TOT	-.1010	1.0162	1.0081	382
PA	-.1578	1.2681	1.1261	214
CA	-.0757	.9131	.9556	263
CAT	-.0368	.7670	.8758	262
GRADE 6, MATH				
TOT	-.1973	.8098	.8999	259
PA	-.2066	.8203	.9057	129
CA	-.1946	.8552	.9248	171
CAT	-.1177	.6840	.8270	184

KEY

TOT = Total Combined Ability Estimate
PA = Paper-Administered Ability Estimate
CA = Computer-Administered Ability Estimate
CAT = Computerized Adaptive Testing Ability Estimate

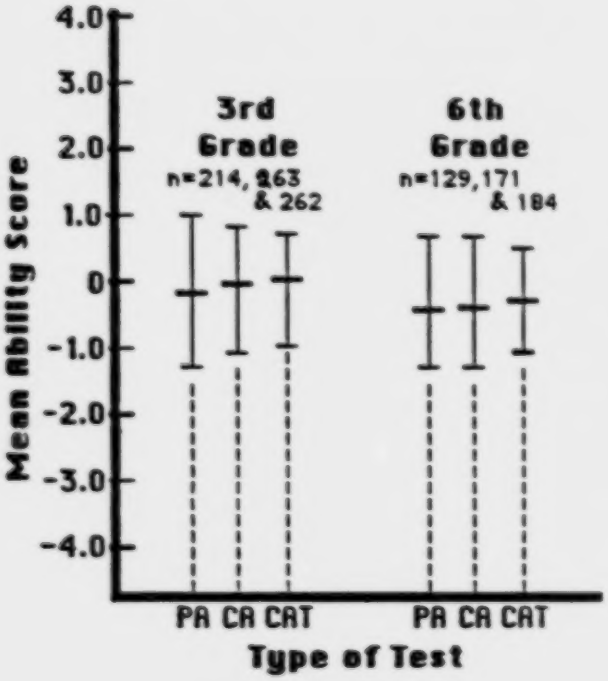
Figure 1 presents a graphic display showing the relation between the ability estimates and testing time required by paper-administered tests (PA), computer-administered tests (CA), and computerized adaptive tests (CAT). This figure shows similar and overlapping distributions of ability estimated with each testing method. In comparison, the CAT tests have a mean that is closer to zero and a smaller standard deviation than the computer-administered or paper-administered tests.

Table 6 presents a pair-wise correlation matrix of the ability estimates for each of the three test administration methods. The correlation matrix shows very high correlations between the total ability estimate and the estimates from paper-administered, computer-administered and computerized adaptive tests. The ability estimated from a paper-administered test is correlated .87 with the computer-administered test, and .81 to .85 with the computerized adaptive test. The ability estimated from a computer-administered test is correlated .83 to .91 with the computerized adaptive test.

Figure 1

Results from CAP Field Study
(Test of Mathematics Applications)

Descriptive Statistics for Ability Distributions and Testing Time in Minutes Required by Type of Test Administration Mode



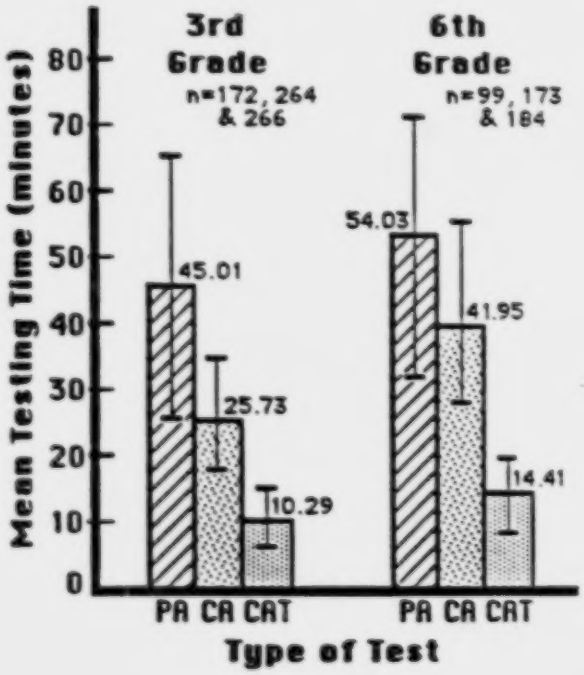


Table 6
Correlation Matrix of Ability Estimates

POSITION	LABEL	1 TOTAL ABILITY	2 PA ABILITY	3 CA ABILITY	4 CAT ABILITY
GRADE 3, MATH					
1	TOTAL ABILITY	100	97	97	92
2	PA ABILITY	97	100	87	85
3	CA ABILITY	97	87	100	83
4	CAT ABILITY	92	85	83	100
GRADE 6, MATH					
1	TOTAL ABILITY	100	97	98	95
2	PA ABILITY	97	100	87	81
3	CA ABILITY	98	87	100	91
4	CAT ABILITY	95	81	91	100

Item Parameters.

Appendix A presents comparisons of the item parameters estimated from the one, two, and three parameter item calibrations for each grade. Table 7 summarizes the distributions of the item parameter estimates from the one, two, and three parameter models. At Grade 3, Table 7 shows an increasingly more normal distribution of item difficulty values as the number of parameters increases. For Grade 6, the two parameter model shows the most normal distribution. Table 7 also shows increasingly higher item discrimination values as the number of parameters increases. The guessing parameters ranged from .06 to .10 for the low value, .17 to .20 for the average value, and .34 to .37 for the high value.

Table 7

ITEM PARAMETER ESTIMATES AND MODELS

	Difficulty			Discrimination			Guessing
PARMS	1	2	3	1	2	3	3
GRADE 3							
MEAN	-.909	-.807	-.618	.737	.860	1.162	.177
S.D.	.808	.831	.992	-	.305	.682	.057
LO	-2.889	-2.969	-3.426		.281	.341	.059
HI	.570	.747	.919		1.712	3.514	.341
GRADE 6							
MEAN	.015	.143	.340	0.561	.643	1.262	.208
S.D.	.716	.896	.653	-	.270	.578	.058
LO	-1.632	-1.579	-1.190		.197	.360	.103
HI	2.199	4.002	2.076		1.627	3.079	.374

The BILOG program calculates a chi-square fit statistic for each item which compares how closely the item characteristic curve fits the actual observed data for each item. The program also provides the degrees of freedom for testing the chi-square statistic and a probability value for the statistic. This chi-square fit statistic allows for comparison of how closely the one, two, and three parameter calibration estimates fit the observed item characteristic curve data. If the chi-square statistic was more than twice as large as the number of degrees of freedom, items were flagged as not fitting the selected item response model. Table 8 presents a summary of the number of items not fitting the one, two, or three parameter models. This table shows that both the two or the three parameter models provide the best fit to the data. Either of these models could be recommended based on this data. The one parameter model did not provide a close fit to the data. For the one parameter model approximately 31% of the Grade 3 items did not fit the model; for grade 6 approximately 18% of the items did not fit the model.

Table 8
Number of Items Not Fitting the IRT Model
Using an Item Chi-Square Fit Statistic

PARMETERS	One	Two	Three	N Items
Grade 3	17	6	7	55
GRADE 6	11	2	2	62

Test Information and Standard Error

Test information and standard error curves were calculated for the one, two, and three parameter models by grade. These test information and standard error curves are presented in Appendix B. These figures show increasing test information, decreasing standard errors, and increasing kurtosis as the number of parameters increases from one to two, and from two to three. Table 9 summarizes the data for point of maximum test information, maximum test information value, and minimum standard error for each grade and each parameter model.

Table 9

Point of Maximum Test Information,
Maximum Test Information Value
and Standard Errors

	Grade 3 Number of Parameters			Grade 6 Number of Parameters		
	One	Two	Three	One	Two	Three
Point of Maximum Test Information	-.86	-.35	.07	.00	-.36	1.07
Test Information Value	17.61	27.07	41.21	12.83	19.72	33.53
Minimum Standard Error	.25	.20	.17	.29	.26	.20

The point of maximum information for the sixth grade, three parameter calibration shows a significant shift to the right on the ability scale indicating higher test information for ability measurements which are one standard deviation above the mean of 0. The point of maximum information for the third grade, one parameter calibration shows a moderate shift to the left on the ability scale indicating (for the one parameter case) higher test information for ability measurements which are .85 standard deviations below the mean of 0. All of the test information curves for one, two, and three parameter models and for both grades are bell-shaped around their respective centroids. The standard error curves for the third

grade tests for all parameter models show smaller standard errors for the lower half of the ability distribution, indicating more accurate measurement for lower ability students than for higher ability students. The standard error curves for the sixth grade test for the one and two parameter models show equivalent standard errors for ability measurements for both low and high ability students. However, the standard error curve for the sixth grade, three parameter calibration shows lower standard errors for the upper half of the ability distribution. This indicates that the sixth grade test, using a three parameter model, is more accurate in measuring higher ability students than for lower ability students. A further result is that the point of maximum test information is substantially higher under the three parameter model than the one or two parameter models indicating that some students are guessing. The presence of empirical guessing data would shift to the right the point of maximum information for each item. The items would be most informative at some higher point than a .50 probability of observed correct responses.

The BILOG program also calculates the log likelihood of the data for each of the parameter models. The log likelihoods and likelihood differences for each grade and model are presented in Table 10.

Table 10
Log Likelihoods and Log Likelihood Differences

	-2 Log Likelihood			Likelihood Differences		
	1P	2P	3P	2P-1P	3P-1P	3P-2P
Grade 3	30788.66	30225.17	29968.18	563.49	820.48	256.99
Grade 6	26427.67	26005.62	25841.78	422.05	585.89	163.84

The results show strong evidence that the three parameter model fits the data significantly better than the two parameter model, and that the two parameter model fits significantly better than the one parameter model.

Score Distribution Tables.

Table 11 summarizes the score distribution tables for each of the ability groups (in increments of .5 from -3.0 to +3.0 theta levels) and testing methods. Comparisons are made for each ability group for the overall ability estimate, the paper-administered test, the computer-administered test, and the computerized adaptive test. The rows indicate the number of students within each cell, the average number right score for the students within each cell, and the proportion of students for each column falling within each cell. The normal distribution probabilities are also presented to the right of the table. An examination of the proportions across a row and within a designated ability group shows that all testing methods produce approximately normal distributions. The majority of proportions are within five percentage points of what would be expected from a normal distribution. These tables also demonstrate that similar ability score distributions are obtained from each of the testing methods.

Table 11
Score Distribution Tables for Ability Groups
and Testing Methods - GRADE 3

ABILITY GROUP		TESTING METHOD OVERALL	PAPER	COMPUTER	CAT	ROW TOTALS	NORMAL DISTRIB.
-3.0 TO -2.5	IN	8	7	2	1	18	
	IS	18.250	11.857	8.000	1.000	13.667	
	IP	.02	.033	.008	.004	.016	.006
-2.5 TO -2.0	IN	13	12	9	4	38	
	IS	28.077	16.250	16.333	5.000	19.132	
	IP	.03	.056	.034	.015	.034	.017
-2.0 TO -1.5	IN	11	6	10	8	35	
	IS	35.364	17.667	18.500	7.750	21.200	
	IP	.028	.028	.038	.031	.031	.044
-1.5 TO -1.0	IN	37	24	29	23	113	
	IS	33.865	22.208	20.621	8.826	22.894	
	IP	.097	.112	.110	.088	.101	.092
-1.0 TO -0.5	IN	57	23	34	36	150	
	IS	39.842	26.522	27.294	10.333	27.873	
	IP	.149	.107	.129	.137	.134	.150
-0.5 TO 0.0	IN	58	38	43	62	201	
	IS	46.466	34.316	33.884	10.694	30.443	
	IP	.152	.178	.164	.237	.179	.192
0.0 TO 0.5	IN	75	38	53	50	216	
	IS	58.827	40.289	40.830	13.500	40.657	
	IP	.196	.178	.202	.191	.193	.192
0.5 TO 1.0	IN	80	32	51	49	212	
	IS	68.425	47.344	47.725	16.510	48.264	
	IP	.209	.150	.194	.187	.189	.150
1.0 TO 1.5	IN	32	25	26	20	103	
	IS	83.625	51.680	51.577	18.200	55.078	
	IP	.084	.117	.099	.076	.092	.092
1.5 TO 2.0	IN	7	6	5	9	27	
	IS	82.571	53.167	53.600	19.778	49.741	
	IP	.018	.028	.019	.034	.024	.044
2.0 TO 2.5	IN	4	3	1		8	
	IS	83.500	55.000	55.000		69.250	
	IP	.010	.014	.004		.007	.017
Total N		382	214	263	262	1121	
Mean		53.908	35.762	36.475	12.775	36.740	

N= Number of Students Per Cell
S= Average Number Right Score
P= Proportion of Total Students Per Cell

Table 11 (Continued)
Score Distribution Tables for Ability Groups
and Testing Methods -GRADE 6

ABILITY GROUP		TESTING METHOD OVERALL	PAPER	COMPUTER	CAT	ROW TOTALS	NORMAL DISTRIB.
-2.5 TO -2.0	N	5	1	1		7	
	S	16.400	14.000	9.000		15.000	
	P	.019	.008	.006		.009	.017
-2.0 TO -1.5	N	19	10	13	9	51	
	S	22.895	13.100	13.615	4.000	15.275	
	P	.073	.078	.101	.050	.069	.044
-1.5 TO -1.0	N	29	22	26	20	97	
	S	24.241	17.955	17.385	6.150	17.247	
	P	.112	.171	.152	.109	.131	.092
-1.0 TO -0.5	N	42	14	26	31	113	
	S	30.690	22.500	21.962	8.516	21.584	
	P	.162	.109	.152	.168	.131	.150
-0.5 TO 0.0	N	47	30	27	40	144	
	S	39.894	28.267	27.667	10.300	26.958	
	P	.181	.233	.158	.217	.194	.192
0.0 TO 0.5	N	56	22	37	41	156	
	S	45.750	36.136	35.378	11.951	33.051	
	P	.216	.171	.216	.223	.210	.192
0.5 TO 1.0	N	33	15	24	30	102	
	S	58.545	43.533	43.708	14.267	39.824	
	P	.127	.116	.140	.163	.137	.150
1.0 TO 1.5	N	25	14	12	10	61	
	S	78.800	51.857	51.250	16.300	56.951	
	P	.097	.109	.070	.054	.082	.092
1.5 TO 2.0	N	2	1	4	2	9	
	S	94.000	57.000	55.500	17.500	55.778	
	P	.008	.008	.023	.011	.012	.044
2.0 TO 2.5	N	1		1	1	3	
	S	78.000		59.000	19.000	52.000	
	P	.004		.006	.005	.004	.004
Total N		259	129	171	184	743	
Mean		42.911	30.496	30.468	10.707	29.917	

N= Number of Students Per Cell
S= Average Number Right Score
P= Proportion of Total Students Per Cell

Correlations of Ability Estimates and Ability Ranks.

Correlation matrices were calculated for ability estimates and ability ranks derived from the one, two, and three parameter models for each of the testing methods. Table 12 presents the correlation matrices for the ability estimates correlated from one, two, and three parameter models and for each method of test presentation. All correlations presented are significant at the .01 level. Since the same data set was used to calibrate the item parameters and then to estimate the ability estimates the following correlations are somewhat higher than would be the case if separate data sets were used for the item calibration and ability estimation.

Ability ranks were calculated by assigning sequential ranks to each ability estimate, from the highest to the lowest, for each of the testing methods and item parameter models. A high correlation would indicate consistent ranking or ordering of students on each of the tests. A high correlation would also indicate that the two tests were equated or that the student would be "indifferent" as to which of the two tests he or she would take, since each of the tests would consistently order the students in the same way. Table 13 presents the correlation matrices for the ability ranks. All correlations presented are significant at the .01 level. The correlations of ability ranks presented in this table are somewhat higher than would be the case if separate data sets were used to estimate item parameters and to calculate ability ranks.

True Score Equating.

True score equating of the paper-administered, computer-administered, and computerized adaptive tests was conducted following the procedures recommended by Lord (1980) and Stocking and Lord (1983). Number right true scores were calculated for each of three ability estimation techniques (Owen's Bayesian, EAP, and maximum likelihood). Correlation matrices were prepared for each of the three true score estimation procedures, for each of three test administration methods, and the proportion correct from each testing method. Table 14 presents the true score equating correlation matrix.

Table 12

SIGNIFICANT CORRELATIONS OF ABILITY ESTIMATES
GRADE 3

POSITION LABEL	1 3TOT	2 3PA	3 3CA	4 3CAT	5 2TOT	6 2PA	7 2CA	8 2CAT	9 1TOT	10 1PA	11 1CA	12 1CAT
1 3TOT	100	97	97	92	96	94	94	90	96	95	94	90
2 3PA	97	100	87	85	94	96	85	84	94	97	85	85
3 3CA	97	87	100	83	96	88	97	82	96	89	97	81
4 3CAT	92	85	83	100	91	84	83	99	91	83	82	99
5 2TOT	96	94	96	91	100	98	98	92	99	97	97	91
6 2PA	94	96	88	84	98	100	90	84	97	99	89	84
7 2CA	94	85	97	83	98	90	100	83	98	90	99	81
8 2CAT	90	84	82	99	92	84	83	100	91	84	82	99
9 1TOT	96	94	96	91	99	97	98	91	100	98	98	92
10 1PA	95	97	89	83	97	99	90	84	98	100	90	84
11 1CA	94	85	97	82	97	89	99	82	98	90	100	81
12 1CAT	90	85	81	99	91	84	81	99	92	84	81	100

SIGNIFICANT CORRELATIONS OF ABILITY ESTIMATES
GRADE 6

POSITION LABEL	1 3TOT	2 3PA	3 3CA	4 3CAT	5 2TOT	6 2PA	7 2CA	8 2CAT	9 1TOT	10 1PA	11 1CA	12 1CAT
1 3TOT	100	97	98	95	97	94	96	94	97	93	96	93
2 3PA	97	100	87	81	95	97	88	80	96	97	89	81
3 3CA	98	87	100	91	96	85	98	90	96	84	98	90
4 3CAT	95	81	91	100	93	80	90	99	92	79	89	99
5 2TOT	97	95	96	93	100	97	98	94	99	96	98	93
6 2PA	94	97	85	80	97	100	89	81	98	99	90	80
7 2CA	96	88	98	90	98	89	100	91	98	88	99	90
8 2CAT	94	80	90	99	94	81	91	100	93	80	90	99
9 1TOT	97	96	96	92	99	98	98	93	100	98	98	93
10 1PA	93	97	84	79	96	99	88	80	98	100	89	80
11 1CA	96	89	98	89	98	90	99	90	98	89	100	89
12 1CAT	93	81	90	99	93	80	90	99	93	80	89	100

KEY
1,2,3 = Number of Parameters Estimated
TOT = Total Ability Estimate
PA = Paper-Administered Test
CA = Computer-Administered Test
CAT = Computerized Adaptive Test

Table 13
SIGNIFICANT CORRELATIONS OF ABILITY RANKS
GRADE 3

POSITION	1	2	3	4	5	6	7	8	9	10	11	12
LABEL	3TOT	3PA	3CA	3CAT	2TOT	2PA	2CA	2CAT	1TOT	1PA	1CA	1CAT
1 3TOT	100	98	98	92	99	97	97	92	99	97	97	91
2 3PA	98	100	91	88	98	99	90	87	97	99	89	87
3 3CA	98	91	100	82	98	90	99	83	97	90	99	80
4 3CAT	92	88	82	100	92	87	82	99	91	86	82	99
5 2TOT	99	98	98	92	100	98	98	92	99	98	97	91
6 2PA	97	99	90	87	98	100	90	87	97	99	89	87
7 2CA	97	90	99	82	98	90	100	82	98	90	99	80
8 2CAT	92	87	83	99	92	87	82	100	91	86	82	99
9 1TOT	99	97	97	91	99	97	98	91	100	98	98	91
10 1PA	97	99	90	86	98	99	90	86	98	100	89	86
11 1CA	97	89	99	82	97	89	99	82	98	89	100	81
12 1CAT	91	87	80	99	91	87	80	99	91	86	81	100

SIGNIFICANT CORRELATIONS OF ABILITY RANKS
GRADE 6

POSITION	1	2	3	4	5	6	7	8	9	10	11	12
LABEL	3TOT	3PA	3CA	3CAT	2TOT	2PA	2CA	2CAT	1TOT	1PA	1CA	1CAT
1 3TOT	100	97	98	94	99	96	98	94	99	95	97	93
2 3PA	97	100	88	80	97	99	88	80	97	98	88	79
3 3CA	98	88	100	91	98	86	99	91	97	84	98	90
4 3CAT	94	80	91	100	94	81	90	99	92	80	89	99
5 2TOT	100	97	98	94	100	97	98	94	99	96	97	93
6 2PA	96	99	86	81	97	100	86	80	97	99	86	80
7 2CA	98	88	99	90	98	86	100	91	97	84	99	90
8 2CAT	94	80	91	99	94	80	91	100	92	80	89	99
9 1TOT	99	97	97	92	99	97	97	92	100	97	98	92
10 1PA	95	98	84	80	96	99	84	80	97	100	85	80
11 1CA	97	88	98	89	97	86	99	89	98	85	100	89
12 1CAT	93	79	90	99	93	80	90	99	92	80	89	100

KEY
1,2,3 = Number of Parameters Estimated
TOT = Total Ability Estimate
PA = Paper-Administered Test
CA = Computer-Administered Test
CAT = Computerized Adaptive Test

Table 14

Grade 3 Correlation Matrix of Number Right and True Score Estimates for Three Ability Estimation Algorithms and Three Testing Methods

	N	PA COR	PA BAY	PA EAP	PA MLE	CA COR	CA BAY	CA EAP	CA MLE	CAT COR	CAT BAY	CAT EAP	CAT MLE
PA-TESTS													
PA- NCorrect	214	100	98	98	98	89	89	89	89	84	83	84	85
PA- BAY True	214	98	100	99	99	88	90	90	90	86	87	88	89
PA- EAP True	214	98	99	100	99	88	89	89	89	87	87	87	88
PA- MLE True	214	98	99	99	100	88	89	89	89	87	86	87	89
CA-TESTS													
CA- NCorrect	261	89	88	88	88	100	98	98	98	82	77	78	80
CA- BAY True	261	89	90	89	89	98	100	99	99	82	80	80	82
CA- EAP True	261	89	90	89	89	98	99	100	99	81	79	80	81
CA- MLE True	261	89	90	89	89	98	99	99	100	82	79	80	82
CAT TESTS													
CAT- NCorrect	262	84	86	87	87	82	82	81	82	100	94	95	97
CAT- BAY True	262	83	87	87	86	77	80	79	79	94	100	99	97
CAT- EAP True	262	84	88	87	87	78	80	80	80	95	99	100	99
CAT- MLE True	262	85	89	88	89	80	82	81	82	97	97	99	100

Table 14 (Continued)

Grade 6 Correlation Matrix of Number Right and True Score Estimates for Three Ability Estimation Algorithms and Three Testing Methods

	N	PA COR	PA BAY	PA EAP	PA MLE	CA COR	CA BAY	CA EAP	CA MLE	CAT COR	CAT BAY	CAT EAP	CAT MLE
PA-TESTS													
PA- NCorrect	129	100	98	98	98	88	88	89	88	77	78	78	76
PA- BAY True	129	98	100	99	99	91	91	92	91	77	78	78	76
PA- EAP True	129	98	99	100	99	90	91	92	91	78	79	79	78
PA- MLE True	129	98	99	99	100	90	90	91	91	78	79	79	78
CA-TESTS													
CA- NCorrect	171	88	91	90	90	100	99	98	99	84	88	88	87
CA- BAY True	171	88	91	91	90	99	100	99	99	84	90	89	88
CA- EAP True	171	89	92	92	91	98	99	100	99	85	90	90	89
CA- MLE True	171	88	91	91	91	99	99	99	100	85	90	90	89
CAT TESTS													
CAT- NCorrect	183	77	77	78	78	84	84	85	85	100	96	97	97
CAT- BAY True	183	78	78	79	79	88	90	90	90	96	100	99	99
CAT- EAP True	183	78	78	79	79	88	89	90	90	97	99	100	99
CAT- MLE True	183	76	76	78	78	87	88	89	89	97	99	99	100

DISCUSSION

This study has demonstrated high levels of comparability in terms of measurement precision between paper-administered tests, computer-administered tests, and computerized adaptive tests. The analysis of variance showed no significant differences between paper-administered tests and computer-administered tests. There was a significant difference for the test order effect demonstrating that student scores were slightly lower on the second test administration when compared to the first test administration. This result was interpreted as a possible test fatigue effect. Counterbalancing of the experimental treatments allowed us to account for this effect. The item analysis results also showed high levels of comparability between the paper-administered tests and the computer-administered tests. The test score means, standard deviations, reliabilities, and standard errors of the means were very similar for the paper-administered and the computer-administered tests.

Ability estimates based on a three parameter IRT model were calculated for each of the three methods of test administration. These results showed that the computerized adaptive tests produced an ability distribution with a mean closer to 0.0 and a smaller variance than either the paper-administered or computer-administered tests.

The computerized adaptive tests also required only one-fourth of the testing time required by the paper-administered tests, while the computer-administered tests required only one-half to three-quarters of the testing time required by the paper-administered tests. The data from this study indicate that the computer-administered tests and the paper-administered tests produce equivalent scores, but with significantly less testing time required for the computer-administered tests. The computerized adaptive tests also provide a more precise ability estimate with smaller variance than the paper-administered or computer-administered tests. In addition, the computerized adaptive tests require significantly less testing time than either the paper-administered tests or the computer-administered tests. The ability estimates calculated from a 20 item computerized adaptive test tend to show more precision than tests of length 55 or 62 items from paper-administered or computer-administered tests.

High correlations were found between the ability estimates from paper-administered, computer-administered, and computerized adaptive tests. The ability estimated from the paper-administered test was correlated .87 with the computer-administered test and .81 to .85 with the computerized adaptive test. The ability estimated from the computer-administered test was correlated .83 to .91 with the computerized adaptive test.

A comparison of calibrations from one, two and three parameter models showed increasing test information, reduced standard errors, and higher kurtosis as the number of parameters increased from one to two, and from two to three. An analysis of the log likelihood differences between the one, two, and three parameter models also showed that the the three parameter model provided a better fit than the two parameter model, and that the two parameter model provided a better fit than the one parameter model. Similar results for each of these effects has been found for the Armed Services Vocational Aptitude Battery subtests (Mislevy and Bock, 1984). They found that

the three parameter model generally provided a better fit to the data than the two parameter model, and that the two parameter model provided a better fit than the one parameter model. They also found that the point of maximum test information was significantly higher for the three parameter model than for the two parameter model. This would indicate that some of the students were occasionally guessing on the test items, and thus the need to account for the pseudo-guessing parameter in the item analysis.

High levels of equivalence were also found between the distributions of ability estimates for paper-administered, computer-administered, and computerized adaptive tests.

High correlations were also found between the ability estimates and ability ranks calculated for each of the one, two, and three parameter models and for the paper-administered tests, computer-administered tests, and computerized adaptive tests. These results indicate that ability estimates could have been calculated from any of the three IRT models and that there was a high intercorrelation among the ability estimates. Furthermore, the high correlations indicate consistent ranking or ordering of the students irrespective of the IRT model chosen or the mode of test presentation.

Several additional studies have also shown close equivalence between paper-administered tests, computer-administered tests and computerized adaptive tests (Kingsbury and Weiss, 1981; Vicino and Hardwicke, 1984; McBride and Martin, 1983; Moreno, Wetzel, McBride and Weiss, 1983; and McKinley and Reckase, 1984; Calvert and Westfall, 1982; Hardwicke and Yoes, n.d.; Hoffman and Lundberg, 1976; Sorenson, 1985). These studies have shown that the scores from paper-administered tests, computer-administered tests, and computerized adaptive tests are very similar. Where there are differences, the computer-administered or computerized adaptive tests are generally slightly easier than the paper-administered tests.

The results from this study indicate that either the two or three parameter IRT models could be recommended for use. The one parameter model did not provide as close a fit to the data.

The study also showed high correlations between the obtained number right score and the estimated number right true scores for three different ability estimation algorithms (Bayesian, EAP, and Maximum likelihood) and for the three modes of test presentation.

Further analysis is planned to investigate the dimensionality of the paper-administered, computer-administered, and computerized adaptive tests.

REFERENCES

Armbruster, B.B., Stevens, R.J. and Rosenshine, B. (1977) Analyzing content coverage and emphasis: A study of three curricula and two tests. Technical Report No. 26. Urbana-Champaign, IL: University of Illinois, Urbana-Champaign, Center for the Study of Reading.

Bejar, I.I. Weiss, D.J. and Kingsbury, G.G. (1977) Calibration of an item pool for the adaptive measurement of achievement. Research Report 77-5. Minneapolis, MN: University of Minnesota, Psychometric Methods Program.

Bock, R.D. and Aitkin, M. (1981) Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. Psychometrika, 46, 443-459.

Bock, R.D. and Mislevy, R.J. (1982) Adaptive EAP estimation of ability in a microcomputer environment. Applied Psychological Measurement, 6(4), 431-444.

Brown, J.M. and Weiss, D.J. (1977) An adaptive testing strategy for achievement test batteries. Research Report 77-6. Minneapolis, MN: University of Minnesota, Psychometric Methods Program.

Calvert, E.J. and Westfall, R.C. (1982) A comparison of conventional and automated administration of Raven's Standard Progressive Matrices. International Journal of Man Machine Studies, 17, 305-310.

Floden, R.E., Porter, A.C., Schmidt, W.H. and Freeman, D.J. (1977) Don't they all measure the same thing? Consequences of selecting standardized tests. Research Series No. 25. East Lansing, MI: Michigan State University, Institute of Research on Teaching.

Freeman, D. Kuhs, T. Knappen, L. and Porter, A. (1979) A closer look at standardized tests. Research Series No. 53. East Lansing, MI: Michigan State University, Institute of Research on Teaching.

Hambleton, R.K. (1983) Applications of item response models to criterion-referenced assessment. Applied Psychological Measurement, 7(1), 33-44.

Hambleton, R.K. (1979) Applications of latent trait theory to the development and use of criterion referenced tests. Paper presented at the Annual Meeting of the American Education Research Association and National Council for Measurement in Education. San Francisco.

Hambleton, R.K. and Rovinelli, R.J. (1983) Assessing the dimensionality of a set of test items. Paper presented at the American Educational Research Association, Montreal.

Hambleton, R.K. and Swaminathan, H. (1985) Item Response Theory: Principles and applications. Boston: Kluwer Nijhoff Publishing.

Hawdwicke, S. B. and Yoes, M.E. (n.d.) Attitudes and performance on computerized and paper and pencil tests. Paper presented at the 2nd Annual Air Force Conference on Technology in Training and Education.

Hoffman, K.L. and Lundberg, G.D. (1976) A comparison of computer-monitored group tests with paper and pencil tests. Educational and Psychological Measurement, 36, 791-809.

House, E.R., Glass, G.V., McLean, L. and Walker, D.F. (1978) No simple answer: Critique of the "Follow-Through" evaluation. Harvard Educational Review. 48,128-160.

Kingsbury, G.G. and Weiss, D.J. (1981) A validity comparison of adaptive and conventional strategies for mastery testing. Research Report 81-3, Psychometric Methods Program, Dept. of Psychology, University of Minnesota, MN.

Kingsbury, G.G. and Weiss, D.J. (1983) A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D.J. Weiss (Ed.) New Horizons in Testing: Latent Trait Theory and Computerized Adaptive Testing. New York: Academic Press.

Lord, F.M. (1980) Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

McBride J.R. (1977) Applications of latent trait theory to criterion referenced testing. In D.J. Weiss (Ed.) Proceedings of the 1977 Computerized Adaptive Testing Conference. Minneapolis, MN: University of Minnesota, Department of Psychology.

McBride, J.R. (1981) Psychometric considerations for the the computerized adaptive testing system. San Diego, CA: Navy Personnel Research and Development Center.

McBride, J.R. and Martin, J.T. (1983) Reliability and validity of adaptive ability tests in a military setting. In D.J. Weiss (Ed.) New Horizons in Testing, New York: Academic Press.

McKinley, R. L. and Reckase, M.D. (1984) Implementing an adaptive testing program in an instructional program environment. Paper presented at the annual meeting of the American Educational Research Association,

Mislevy, R.J. and Bock, R.D. (1982) BILOG: Maximum likelihood item analysis and test scoring with logistic models for binary items. Mooresville, ID: Scientific Software.

Mislevy, R. J. and Bock, R.D. (1984) Item operating characteristics of the Armed Services Vocational Aptitude Battery (ASVAB), Form 8A. Final Report Office of Naval Research, Contract No. N00014-83-C-0283.

Moreno, K. Wetzel, C.D., McBride, J.R. and Weiss, D.J. (1983) Relationships between corresponding Armed Service Vocational Aptitude Battery (ASVAB) and computerized adaptive testing (CAT) subtests. (NPRDC TR 83-27) San Deigo: Navy Personnel Research and Development Center,

Olsen, J.B., Bunderson, C.V., and Gilstrap, R.M. (1982) Development of a preliminary design for a computerized adaptive testing service for the Department of Defense. Orem, UT: WICAT Systems, Inc.

Olsen,J.B., Bunderson, C.V., and Gilstrap, R.M. (1983) Prototype design and development of a computerized adaptive testing system. Report for the Department of Defense, U.S. Marine Corps Executive Agent. Orem, UT: WICAT Systems, Inc.

Reckase, M.D. (1985) The difficulty of test items that measure more than one ability. Paper presented at the American Educational Research Association Meeting, Chicago.

Samejima, F. (1974) Normal ogive model on the continuous response level in the multi-dimensional latent space. Psychometrika, 39, 111-121.

Schnitz, J., Maynes, D. and Olsen, J.B. (1983) Evaluation of a schema-based reading comprehension curriculum administered by computer. Report to the U.S. Department of Education. Provo, UT: WICAT Education Institute.

Sorenson, H.B. (1985) Cognitive ability tests. The Monitor. November/December 1985.

Stocking, M.L. (1984) Two simulated feasibility studies in computerized adaptive testing. Paper presented at the American Educational Research Association, New Orleans.

Stocking, M. L. and Lord F.M. (1983) Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201-210.

Traub, R.E. and Wolf, R.G. (1982) Latent trait theories and the assessment of educational achievement. Review of Research in Education,9,377-433.

Tyler, R.W. and White, S.H.(1979) Testing, teaching, and Learning. U.S. Department of Health, Education and Welfare and National Institute of Education.

Vicino, F.L. and Barwicke, S.B. (1984) An evaluation of the utility of computerized adaptive testing. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, 1984.

Watts, K., Baddeley, A. and Williams, M. (1982) Automated tailored testing using Raven's matrices and the Minl Hill Vocabulary Tests: A comparison with manual administration. International Journal of Man Machine Studies, 17, 331-344.

Weiss, D.J. (1975) Computerized adaptive ability measurement. Naval Research Reviews, 28, 1-28.

Weiss, D.J. (1979) Computerized adaptive achievement testing. In H.F. O'Neil (Ed.) Procedures for Instructional Systems Development. New York: Academic Press.

Weiss, D.J. (1981) Computerized adaptive performance evaluation, Final Report. Minneapolis,MN: Computerized Adaptive Testing Laboratory, University of Minnesota.

Weiss, D.J. (Ed.) (1983) New Horizons in Testing: Latent Trait Theory and Computerized Adaptive Testing. New York: Academic Press.

Weiss, D.J. and Kingsbury, G.G. (1984) Application of computerized adaptive testing to educational problems. 21(4), 361-375.

Whitely, S.E. (1980) Multicomponent latent trait models for ability tests. Psychometrika, 45, 479-494.

APPENDIX A

COMPARISONS OF ITEM PARAMETERS FOR ONE, TWO, AND THREE PARAMETER MODELS

APPENDIX A

GRADE 3 MATHEMATICS
ITEM PARAMETERS

PARMS	Difficulty			Discrimination			Guessing
MODEL	1	2	3	1	2	3	3
ITEM #							
1	-0.154	-0.043	0.303	0.737	0.987	1.872	.191
2	0.289	0.318	0.475	"	0.964	1.659	.122
3	-0.916	-0.631	-0.067	"	1.253	3.514	.316
4	-0.603	-0.378	-0.028	etc.	1.279	2.055	.205
5	0.169	0.213	0.390		0.992	1.654	.124
6	-1.379	-1.230	-1.191		0.745	0.685	.145
7	-1.249	-1.299	-1.073		0.561	0.587	.165
8	-1.704	-1.368	-1.319		0.911	0.771	.226
9	-0.701	-0.669	-0.024		0.610	0.860	.284
10	-1.238	-1.256	-1.118		0.580	0.569	.157
11	-1.097	-0.984	-0.939		0.717	0.672	.112
12	-0.041	0.014	0.883		0.409	1.156	.341
13	-1.174	-1.174	-1.065		0.599	0.605	.134
14	-2.037	-1.888	-2.050		0.724	0.615	.180
15	-1.588	-1.259	-1.209		0.932	0.894	.180
16	-1.515	-1.328	-1.300		0.762	0.703	.157
17	-0.948	-0.790	-0.432		0.767	0.883	.219
18	-1.142	-0.958	-0.737		0.796	0.830	.181
19	-0.004	0.074	0.402		0.872	1.985	.192
20	-1.200	-1.005	-0.910		0.810	0.784	.145
21	-2.031	-1.796	-1.876		0.777	0.657	.222
22	-0.863	-0.635	-0.230		1.009	1.394	.241
23	-1.298	-1.010	-1.005		0.945	0.859	.118
24	0.035	0.116	0.439		0.810	1.713	.190
25	-1.002	-0.776	-0.701		0.918	0.905	.111
26	-2.179	-1.600	-1.768		1.113	0.942	.188
27	-0.566	-0.349	-0.036		1.248	1.979	.180
28	-0.040	0.042	0.205		1.277	1.919	.101
29	-1.933	-2.313	-2.318		0.504	0.452	.199
30	-0.392	-0.255	0.032		0.886	1.170	.164
31	-1.737	-1.796	-1.692		0.596	0.551	.211
32	0.570	0.747	0.919		0.559	1.455	.182
33	-2.122	-1.935	-2.180		0.746	0.593	.198
34	-0.793	-0.512	-0.152		1.588	2.845	.220
35	0.245	0.307	0.430		0.807	1.151	.098
36	-0.209	-0.072	0.033		1.712	2.140	.059
37	-2.068	-2.969	-2.807		0.411	0.392	.209
38	-0.857	-0.827	-0.079		0.559	0.771	.307
39	-0.431	-0.254	0.057		1.283	2.290	.177
40	-1.179	-1.514	-1.540		0.771	0.698	.168

APPENDIX A

GRADE 3 MATHEMATICS

PARMS	Difficulty			Discrimination			Guessing
MODEL	1	2	3	1	2	3	3
ITEM #							
41	-0.150	-0.050	0.101		1.120	1.437	.097
42	-1.277	-1.273	-1.025		0.610	0.624	.188
43	-0.581	-0.393	-0.139		1.462	1.874	.148
44	0.363	0.366	0.475		1.068	1.873	.097
45	-2.283	-2.051	-2.347		0.768	0.621	.189
46	0.084	0.140	0.263		1.206	1.668	.080
47	-1.934	-1.904	-1.862		0.656	0.601	.216
48	-2.889	-2.788	-3.426		0.704	0.523	.213
49	0.055	0.176	0.633		0.423	0.564	.177
50	-1.211	-1.006	-0.913		0.820	0.788	.150
51	-0.728	-0.749	-0.286		0.485	0.532	.203
52	-0.762	-0.688	-0.236		0.621	0.752	.223
53	-0.369	-0.243	0.031		1.096	1.636	.160
54	-0.543	-0.843	-0.195		0.281	0.341	.168
55	-0.125	-0.015	0.196		1.169	1.861	.124
MEAN	-0.909	-.807	-.618	.737	.860	1.162	.177
S.D.	.808	.831	.992	-	.305	.682	.057
LO	-2.889	-2.969	-3.426		.281	.341	.059
HI	.570	.747	.919		1.712	3.514	.341

APPENDIX A

GRADE 6 MATHEMATICS
ITEM PARAMETERS

PARMS	Difficulty			Discrimination			Guessing
MODEL	1	2	3	1	2	3	3
ITEM #							
1	-0.023	-0.026	0.344	0.561	0.660	1.118	.230
2	2.199	2.686	1.455	"	0.430	2.173	.103
3	0.827	1.026	1.125	"	0.424	3.079	.267
4	0.294	0.598	1.253	etc.	0.254	0.360	.194
5	-1.632	-1.242	-1.190		0.848	0.824	.192
6	-0.009	-0.010	0.445		0.567	1.093	.257
7	-0.515	-0.352	-0.282		1.003	1.127	.116
8	-0.739	-0.574	-0.200		1.011	1.673	.276
9	0.780	1.290	1.219		0.287	1.594	.268
10	-0.466	-0.337	-0.203		1.016	1.211	.146
11	-0.564	-0.489	-0.264		0.755	0.945	.180
12	0.314	0.356	0.616		0.477	0.754	.180
13	-0.361	-0.319	0.188		0.767	1.532	.297
14	0.138	0.121	0.581		0.620	2.515	.307
15	0.792	0.735	0.757		0.609	1.138	.143
16	0.035	0.013	0.253		0.690	1.034	.180
17	0.652	0.792	1.000		0.423	1.096	.228
18	-0.945	-0.945	-0.734		0.543	0.577	.171
19	0.577	0.487	0.570		0.739	1.418	.155
20	1.537	1.734	1.260		0.475	1.478	.134
21	0.526	0.742	1.078		0.336	2.458	.317
22	-0.658	-0.658	-0.114		0.533	0.696	.259
23	0.583	0.618	0.804		0.498	0.883	.180
24	-0.217	-0.208	0.346		0.541	0.994	.273
25	-1.054	-1.579	-0.992		0.319	0.364	.202
26	0.270	0.203	0.414		0.696	1.152	.174
27	0.018	0.079	0.420		0.353	0.494	.165
28	-0.302	-0.236	-0.009		1.139	1.813	.193
29	0.435	0.536	0.806		0.454	0.793	.193
30	-0.970	-0.722	-0.642		1.032	1.059	.153
31	1.631	4.002	2.076		0.197	1.148	.181
32	-0.282	-0.207	-0.069		1.263	1.703	.142
33	-0.139	-0.141	0.137		0.547	0.678	.169
34	-0.240	-0.196	0.007		0.716	0.931	.163
35	-1.245	-0.767	-0.740		1.627	1.609	.153
36	-1.443	-1.175	-0.922		0.754	0.799	.242
37	-0.656	-0.488	-0.211		1.000	1.393	.220
38	0.324	0.777	1.296		0.221	2.542	.374
39	-0.385	0.354	-0.083		0.687	0.853	.186
40	0.583	0.858	1.110		0.355	2.558	.301

APPENDIX A

GRADE 6 MATHEMATICS
ITEM PARAMETERS

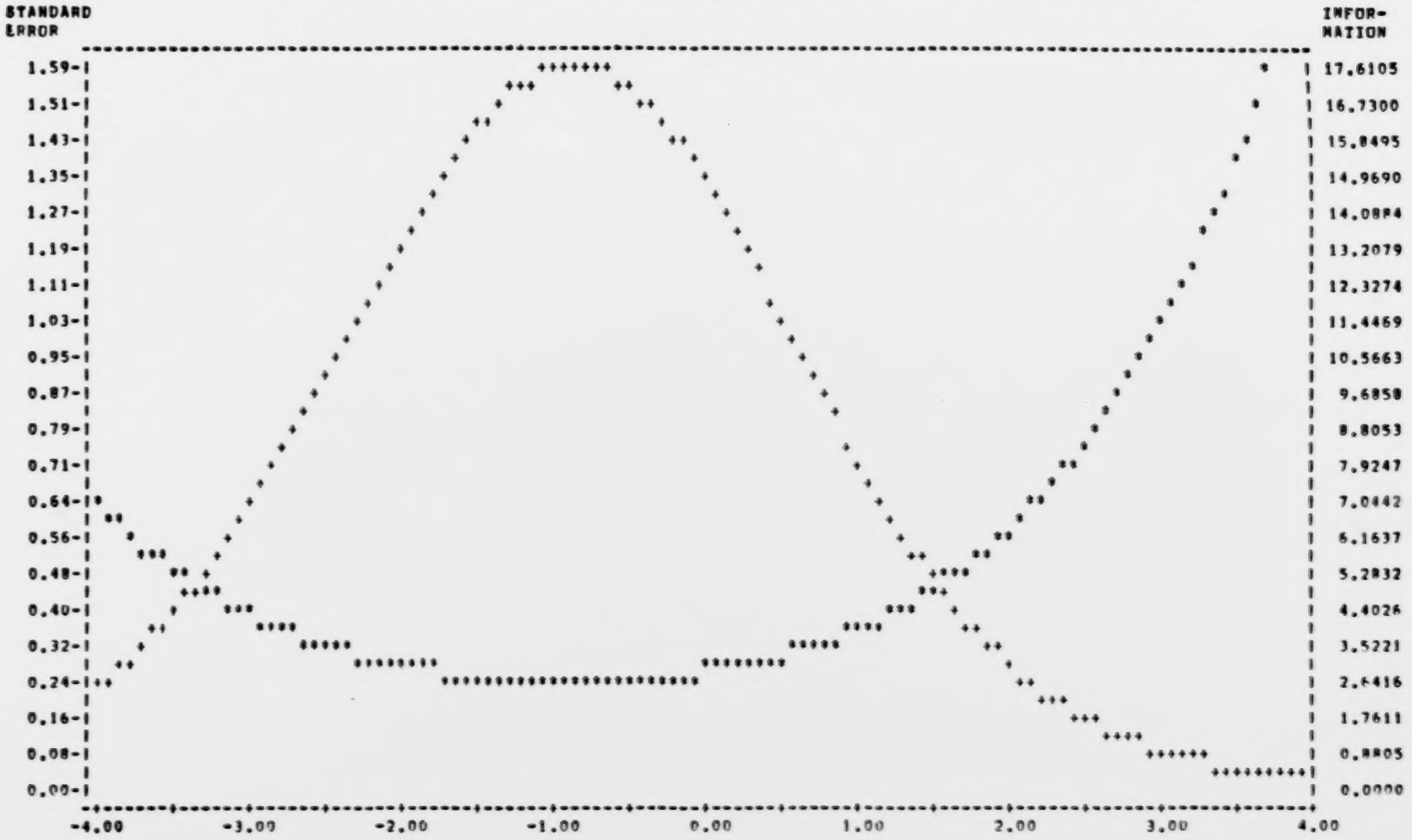
PARMS	Difficulty			Discrimination			Guessing
MODEL	1	2	3	1	2	3	3
ITEM #							
41	0.673	0.560	0.704		0.724	1.978	.190
42	1.241	2.034	1.387		0.296	1.773	.214
43	0.812	0.935	1.019		0.461	0.918	.174
44	-0.155	-0.150	0.126		0.702	0.992	.191
45	-0.060	-0.057	0.066		0.853	1.096	.135
46	-0.349	-0.282	-0.119		0.949	1.201	.154
47	0.062	0.052	0.461		0.674	1.590	.260
48	0.095	0.098	0.300		0.523	0.743	.155
49	0.067	0.077	0.534		0.529	1.130	.262
50	-0.854	-0.866	-0.310		0.523	0.687	.265
51	0.180	0.151	0.493		0.643	1.610	.243
52	-0.076	-0.066	0.362		0.480	0.694	.211
53	0.306	0.327	0.767		0.489	1.759	.290
54	-0.499	-0.424	-0.174		1.007	1.451	.205
55	-0.683	-0.611	-0.456		0.649	0.736	.140
56	-0.235	-0.198	0.167		0.779	1.322	.238
57	0.094	0.129	0.651		0.474	0.906	.256
58	0.088	0.161	0.630		0.321	0.469	.188
59	0.386	0.334	0.549		0.677	1.392	.194
60	-0.389	-0.347	0.198		0.764	1.762	.319
61	0.620	0.477	0.518		0.834	1.330	.120
62	-0.065	-0.071	0.315		0.625	1.026	.226
MEAN	.015	.143	.340	0.561	.643	1.262	.208
S.D.	.716	.896	.653	-	.270	.578	.058
LO	-1.632	-1.579	-1.190		.197	.360	.103
HI	2.199	4.002	2.076		1.627	3.079	.374

APPENDIX B

TEST INFORMATION AND STANDARD ERROR CURVES FOR ONE, TWO, AND THREE PARAMETER MODELS

TEST: E2G3X

TEST INFORMATION AND STANDARD ERROR CURVES FOR THIRD GRADE: ONE PARAMETER MODEL



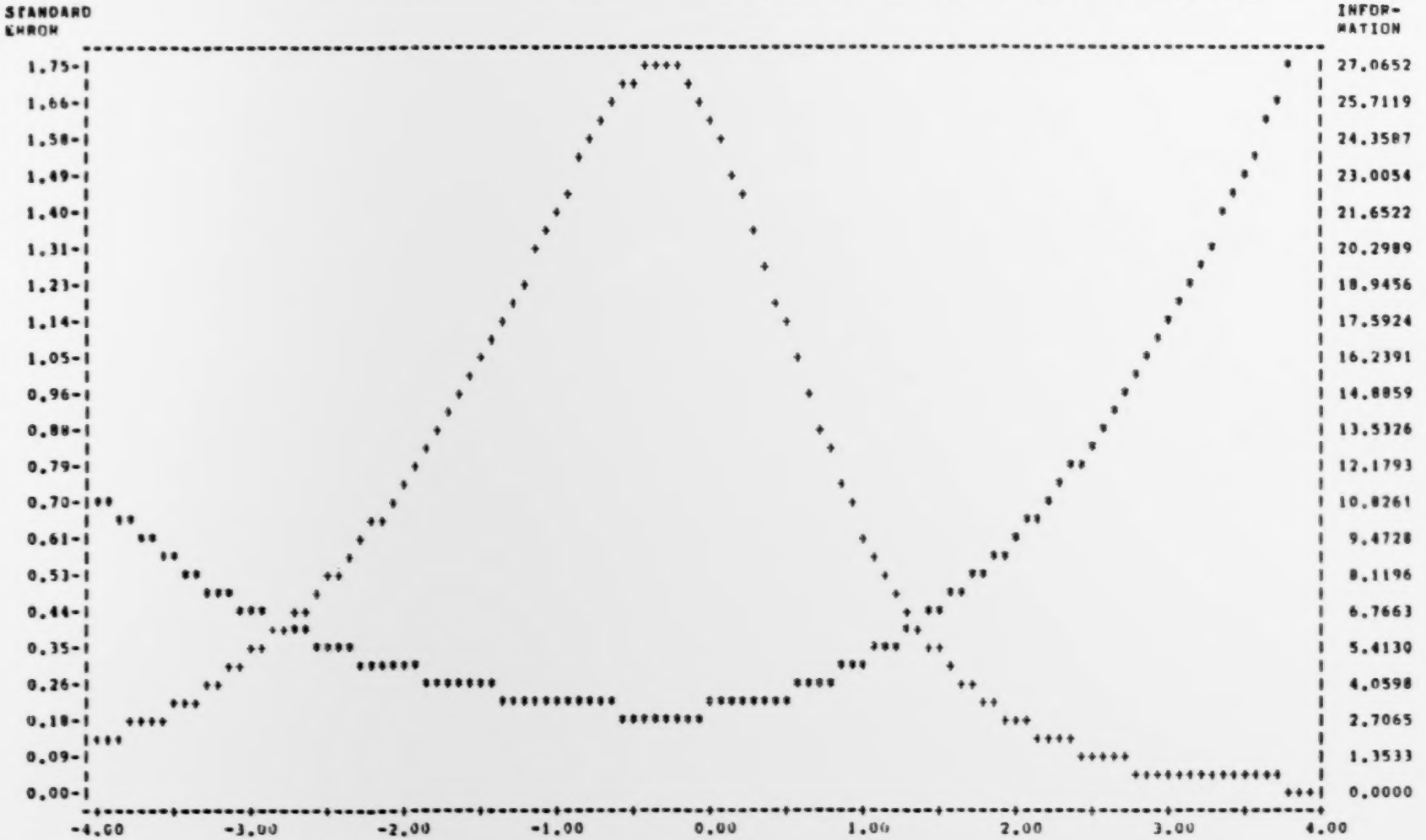
MAXIMUM INFORMATION APPROXIMATELY 0.1761D+02 AT -0.8571

53676 BYTES OF WORKSPACE USED OF 640000 AVAILABLE

BEST COPY AVAILABLE

TEST: B2G3X

TEST INFORMATION AND STANDARD ERROR CURVES FOR THIRD GRADE: TWO PARAMETER MODEL

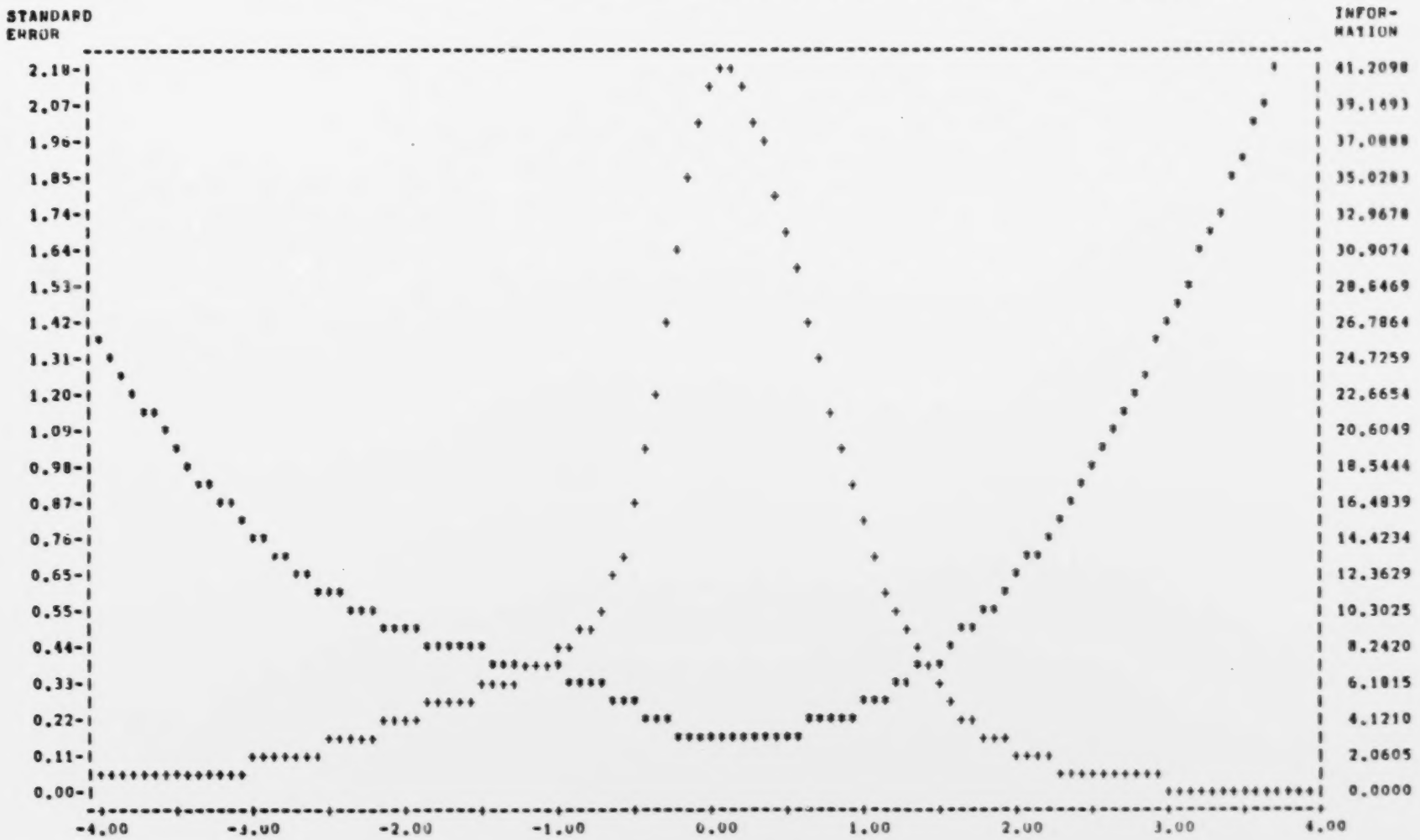


MAXIMUM INFORMATION APPROXIMATELY 0.2707D+02 AT -0.3571

51564 BYTES OF WORKSPACE USED OF 640000 AVAILABLE

BEST COPY AVAILABLE

TEST: E2G3X TEST INFORMATION AND STANDARD ERROR CURVES FOR THIRD GRADE: THREE PARAMETER MODEL

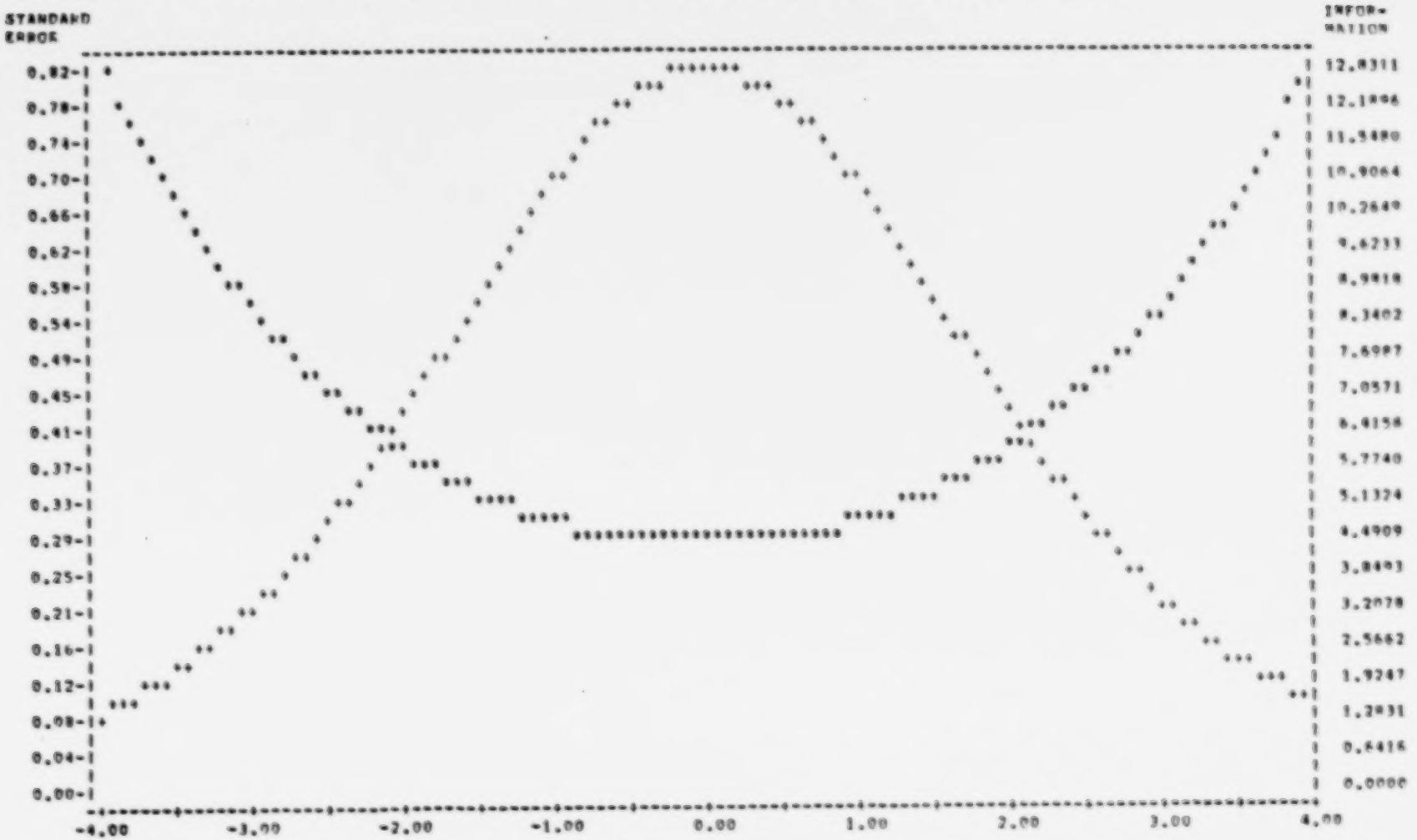


MAXIMUM INFORMATION APPROXIMATELY 0.4121D+02 AT 0.0714

56340 BYTES OF WORKSPACE USED OF 640000 AVAILABLE

BEST COPY AVAILABLE

TEST: E2G6X TEST INFORMATION AND STANDARD ERROR CURVES FOR SIXTH GRADE: ONE PARAMETER MODEL

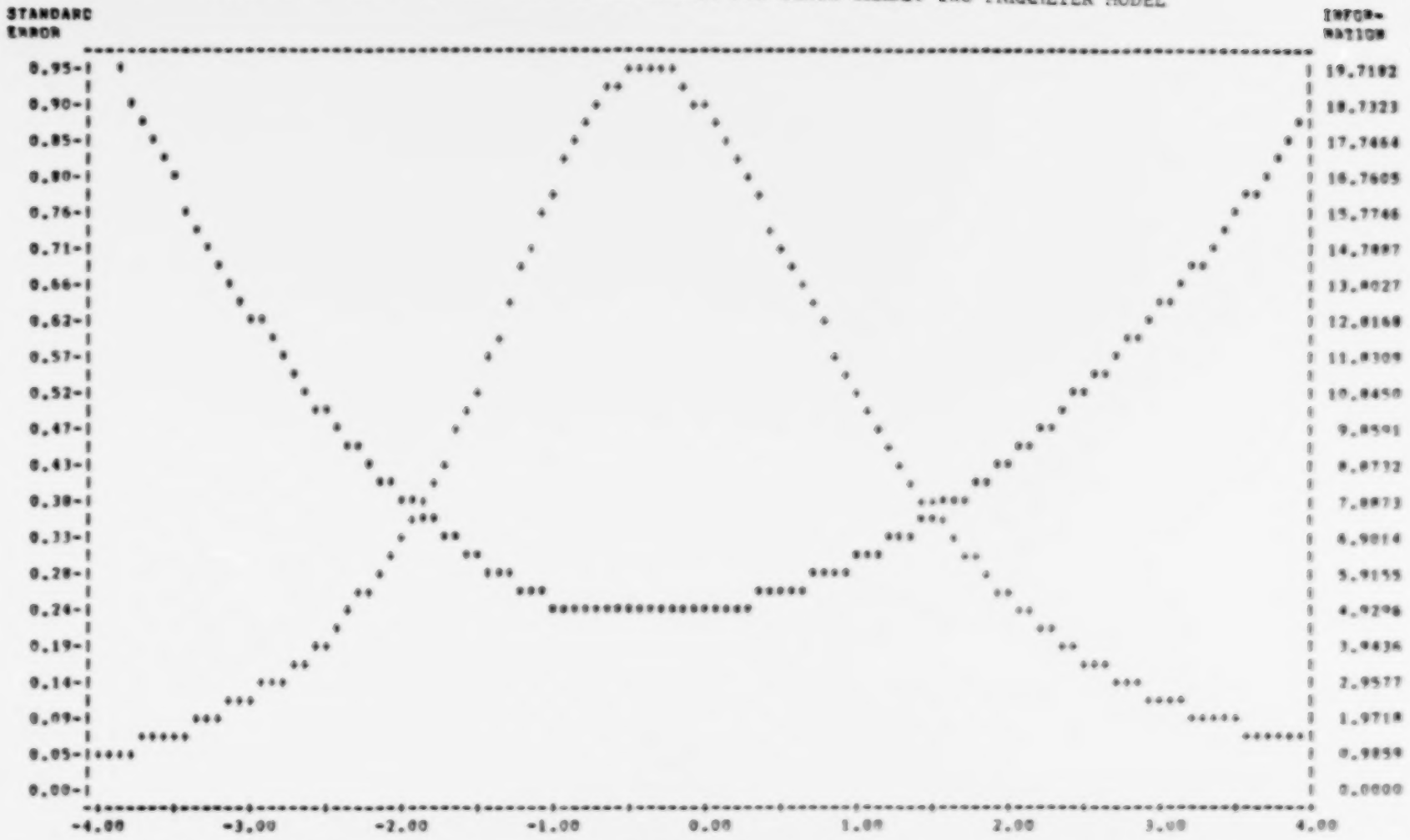


MAXIMUM INFORMATION APPROXIMATELY 0.12830+02 AT 0.0000

60340 BYTES OF WORKSPACE USED OF 640000 AVAILABLE

BEST COPY AVAILABLE

TEST: E2G61 TEST INFORMATION AND STANDARD ERROR CURVES FOR SIXTH GRADE: TWO PARAMETER MODEL

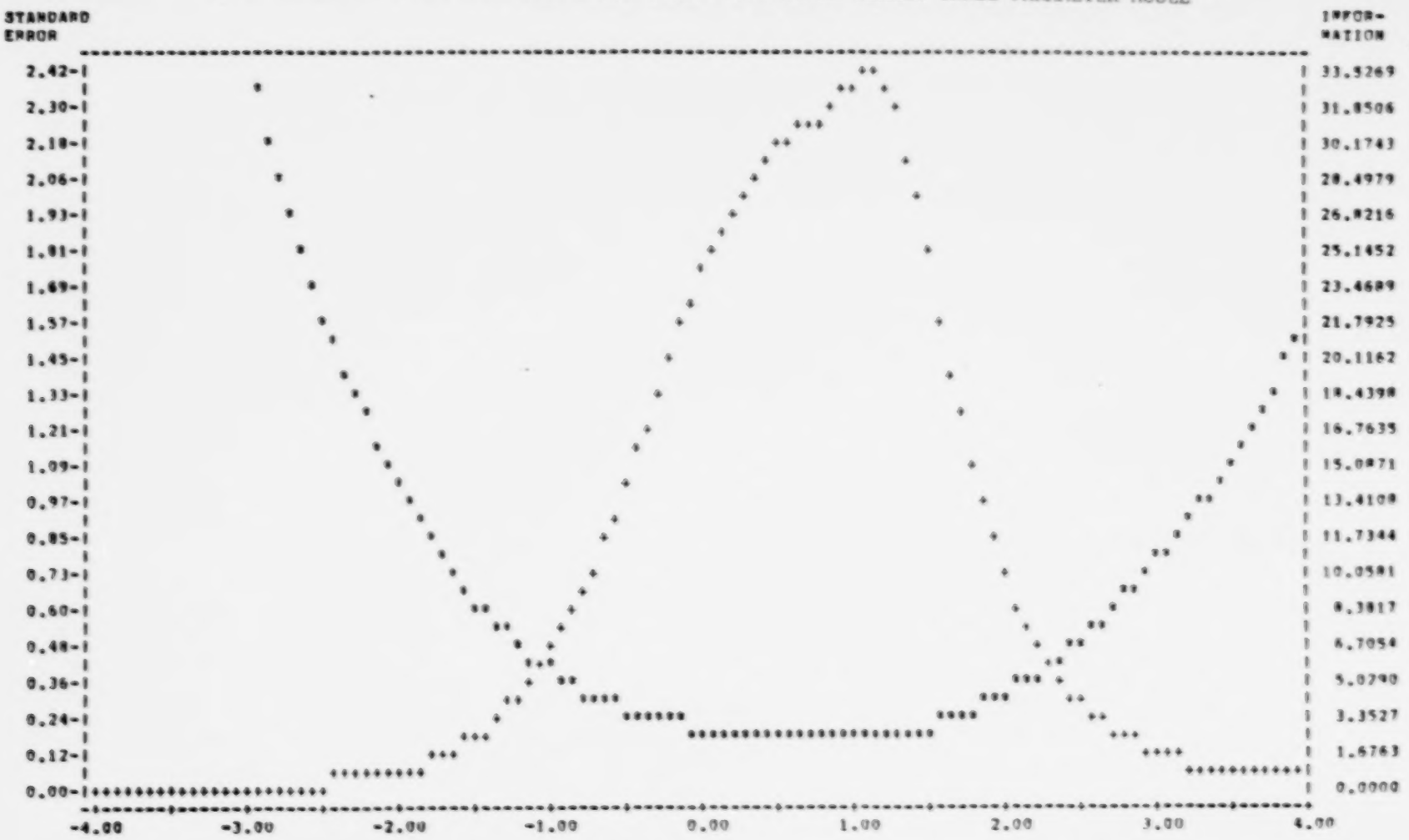


MAXIMUM INFORMATION APPROXIMATELY 0.1972D+02 AT -0.3571

61340 BYTES OF WORKSPACE USED OF 640000 AVAILABLE

BEST COPY AVAILABLE

TEST: E2G6X TEST INFORMATION AND STANDARD ERROR CURVES FOR SIXTH GRADE: THREE PARAMETER MODEL



MAXIMUM INFORMATION APPROXIMATELY 0.33530+02 AT 1.0714

63340 BYTES OF WORKSPACE USED OF 640000 AVAILABLE

BEST COPY AVAILABLE